CAM extended integrations
Example : three years or 3x365 = 1095 days

Step One: Obtaining Inititialization Data

The sources of data are a real problem when more than a limited time period is needed. The only sources that extend for years with a uniform system are the reanalyses. The ERA-40 only goes up to 200. NCEP's reanalysis assimilation system is not good in the Tropics and is not clear how to obtain model level NCEP data. JRA has limited resolution, and moisture only goes to about 100 hPa but it does supply contemporary data. The GDAS data from NOMADS is uncertain as to availability. The problem with GDAS is the archiving of model level data. It is possible to pick up the NCEP GDAS (if we could read it) in real time, this is what NCAR does. This only addresses needs starting now, archiving is not in NCEP's interest.

If the long integration is to be executed at 2° x 2.5° resolution then the initial data horizontal resolution is less of an issue. The vertical structure of all the initial data sources is compatible up to the 30 level CAM configuration needed if the UW PBL/Convection scheme is used.

With the JRA-25 data presently in place, a three year integration could be done in less than a month, end to end. JRA-25 covers from 1979 to present. The strategy would be to wait on the new reanalyses (NASA, ECMWF, NCEP) for higher resolution runs.

Possible Sources

GDAS (0.5° x 0.5°)L63 These data were obtained from the NOMADS server. We have 2004 covering the MPACE and 2006 from Jan to 5 Oct (when Nomads died). These data have problems of availability and decoding. The data format changed and it is not clear that we could decode the new data, if we could get it. I have had problems trying to get a definitive answer on the future of model level data.

NCAR GFS Varying horizontal resolutions and time coverage:
High resolution 0.3125° x 0.3125° / 64 levels -very limited time, part of 2007 availability.
T62 (144x73, 2.5° x 2.5°), available for 2004,2005,2006, 2007 with some holes.
T170 (512 x 256, 0.7° x 0.7° is available for 2006, 2007 with holes.

ERA-Reanalysis (1.25° x 1.25°) 60 Levels - 1979-2001

JRA T106L40 (320 x 160, 1.25° x 1.25°) - easily obtainable, compressed GRIB - no moisture above 100 hPa. The period covered is from 1979 to present (a lag to a few days from real time).

MERRA NASA GEOS5 (1/2° x 2/3°) 72 levels - not available until later in 2008 and even then not certain as to what time periods will be done first. The proposed data distribution would make the data very easy to obtain. The published data plan indicates that the project will provide the data we need.

Other sources ???

Example 1: obtaining NCAR GFS

For the purposed of illustration, an example using the NCAR GFS data will be worked through. This data is currently available on the NCAR MSS.
NCAR MSS ⇒ local NCAR machine ⇒ bassi/NERSC ⇒ llnl/local
time for one (686MB) file :
NCAR MSS ⇒ local NCAR machine: 17 min to 6 min varies on a lot of factors (multiple xfers may have less latency)
local NCAR ⇒ NERSC: 6min
NERSC ⇒ LLNL: 2min 30s
Total = (10 + 6 + 2)*1095*4 = 78840Min = 1314.0h = 54.75 days

The very long time is mostly a feature of the NCAR MSS latency in this test - there are ways to address this problem with some staging techniques on the NCAR machine. The transfer to LLNL is a harder limit. One possible speedup is to just do the one step transfer to NERSC and do the storage and processing on those machines. The mechanism to transfer the data to the desired CAM grid would have to be moved up to NERSC. This would mean that the forcing files would still have to be moved to LC if the integrations are to be done there. In the old days the forcing data was generally much smaller but this may no longer be true as we go to 1° x 1° and 0.5° x 0.5° CAM resolutions.

Example 2: obtaining JRA 25

These data are much smaller than the NCAR/GFS data since they are on a T106L40 grid and in GRIB (compressed) , approximately 2.5 Gb/month. For a 2° x 2.5° L30 CAM integration the JRA resolution is adequate. The JRA does not provide moisture above 100 hPa. This can be addressed in a number of ways. The easiest is to assume that the RH is constant above the last level for reported moisture and perform the calculation using the pressure and temperature. The requisite data for an extended integration can be downloaded in a number of days since the Japanese site is designed for efficient access and transfer.

JRA 25 ⇒ LLNL: 2 days

Put data on CAM grid

Depends on (1) platform, (2) input data size, (3) output data size. For ECMWF TWP-ICE (60 days) on krait 5h 5h*(1095/60) = 90h = 4 days

Spinup

This depends on tradeoffs for number of processors vs. time in queue
Latest LC using 3.5-23 0.9° x 1.25°
full TWP-ICE 32 days = 7.5h which implies
(1095/32)*7.5 =255.0h
number of jobs: 255/12. = 21.5 (limit of 12h per job)
Timing depends on the job queue bottle neck since this is sequential. If you reduce the number

of processors the job spends less time in the queue, but the job takes longer. With Art's new modifications and changing the sequence of loading the data to storage, this time can be smaller. The cruel truth of modern computing is that we can produce numbers much, much faster than we can move them.
Might take a month, most likely quite a bit less.

Production

These jobs can run in parallel - goes much faster than spinup 5h / 10 days depends on processors / job queue - more processors longer in the queue
5*(1095/10) = 545h
Again, this will probably be (substantially ?) faster.

LC / NERSC $\Rightarrow$ local storage

This is for xferring the full files - we could subset on the big machine and only bring down smaller files
LC $\Rightarrow$ localRAID : TWP ICE = 0.9° x 1.25° , all day 8h (1095/60)*8h = 144h